FDG PET Response Monitoring in Malignant Lymphoma

Zijlstra-Baalbergen, J.M.

2008

document version
Publisher's PDF, also known as Version of record

Chapter 4

FDG PET in lymphoma, the need for standardization of interpretation: an observer variation study

Josée M. Zijlstra, Emile F. Comans, Arthur van Lingen, Otto S. Hoekstra, Chad M. Gundy, Jan Willem Coebergh, Vivian Bongers

ABSTRACT

Objective: To measure and describe patterns of interobserver variation in visual interpretation of FDG PET in malignant lymphoma.

Methods: Eleven nuclear medicine physicians with different levels of PET experience independently reviewed 37 FDG PET scans of lymphoma patients (10 obtained at presentation, 27 during or after therapy). They were requested to identify and localize suspicious lymphoma sites and to assign a stage to the baseline scans, and to interpret the remaining scans for the presence of viable lymphoma. Individual (extra-)nodal regions were assessed for likelihood of malignancy as positive, negative or equivocal. These results were compared to expert readings after dichotomisation in conservative and sensitive reading classifications.

Results: 61% and 56% (using sensitive and conservative reading, respectively) of the baseline scans were scored in accordance with the experts. Fourteen of the 27 scans obtained for therapy evaluation with viable tumor sites were scored in accordance with the experts in 82% and 94% of the patients, using conservative and sensitive reading, respectively. The 13 negative scans were scored in agreement with the experts in only 45% of the cases. False positivity pertained especially to the neck, periclavicular, axilla, mediastinum, lung and bone marrow. More experienced observers tended to have less false negative scores.

Conclusion: There are substantial disparities among nuclear medicine physicians' interpretations of FDG PET scans of lymphoma patients, which may affect patient care and results of multi-institutional clinical trials. A well-defined set of criteria is urgently needed to improve consistency.

INTRODUCTION

FDG PET has emerged as a powerful functional imaging tool in treatment individualization of aggressive non-Hodgkin lymphoma (NHL) and Hodgkin's disease (HD) [1;2]. At diagnosis, FDG PET improves the assessment of disease burden.[3;4] In evaluation of therapy, FDG PET differentiates viable tumor from fibrosis in residual masses after chemotherapy. Furthermore, several studies have suggested its use during therapy to identify patients who might benefit from a timely switch to alternative line therapy.[5;6] Although the perspective to implement FDG PET in malignant lymphoma is excellent, some methodological issues remain. PET positivity is based on visual interpretation. However, criteria for test positivity are not uniform in the literature.[7] Moreover, data on observer variation of PET are quite limited. We decided to measure the observer variation, and to identify patterns of error in reading FDG PET scans obtained for staging and therapy evaluation in Hodgkin's and non-Hodgkin's lymphoma.

MATERIALS AND METHODS

Design

Thirty-seven PET scans of lymphoma patients were evaluated on the basis of visual inspection by 11 experienced board-certified nuclear medicine physicians who had a variable experience with PET: five had interpreteted between 750-1250 PET scans in their own practice (further indicated as 'the e-group'); the remaining six had no personal PET experience ('the i-group'). All had taken PET courses in the institute of the expert nuclear medicine physicians and CME courses of the SNM and EANM. Their results were compared with the combined judgement of two expert nuclear medicine physicians (EC, OSH), both with over a decade of PET experience, reviewing > 8000 PET scans. The readings of these experts were used as the gold standard. The whole body PET scans were randomly chosen from the PET referral database of the VUmc (VU University Medical Centre, Amsterdam, The Netherlands), and from prospective studies.[8] Ten scans had been obtained in newly diagnosed HD (n=7) and NHL patients (n=3), at presentation. Fifteen scans were midtreatment scans (4 HD, 11 NHL) for monitoring therapy response, and the remaining 12 scans had been performed to evaluate completed first-line chemotherapy (3 HD, 9 NHL).

PET scans had been performed using a full ring BGO scanner (ECAT EXACT HR$^+$, CTI/Siemens), in 2D mode, with emission scans of 5 min/bed, starting 60 min after 370 MBq FDG administration, and typically covering a scan trajectory of mid-femur to skull. The scans were corrected for decay, scatter, and randoms, and reconstructed using ordered

subset expectation maximization (OSEM) with two iterations and 16 subsets followed by post-smoothing (Hanning 0,5 filter, transaxial spatial resolution 7 mm fool-width of half maximum).

The observers were provided with the same clinical information accompanying the original PET scan referral, but they had no access to other imaging tests, like CT scans (ie. results of physical examination / CT scans at presentation).

The observers were requested to localize and assess the likelihood of malignancy of any potential lymphoma lesion. The readers used a data sheet to score involvement of 9 nodal sites: neck (right and left side), periclavicular (right and left side), axillary (right and left side), mediastinal and hilar (right and left side), para-aortic (abdominal) and iliac, mesenteric and iguinal (right and left side), as well as five extranodal sites: lung, pleura/pericard (termed "serous" in the analyses), liver, spleen and bone marrow. Abnormalities (vs. normal ^{18}F-FDG biodistribution) were asked to be classified as positive, negative or equivocal for the presence of viable lymphoma. Each observer independently interpreted the set of scans using a specially designed software tool (running in Matlab 5.3), which allowed simultaneous visualisation of PET images in the axial, coronal and sagittal planes (at 5 or 10 mm slice thickness), with possible cross linking. This software tool had been installed on the personal computer of each observer, and the results were electronically stored for analysis. In order to be able to accurately relate results of different observers, the coordinates of each hot spot identified by any observer were electronically linked to the assigned interpretation. All observers had worked with this software before during a comparable interobserver study in non-small cell lung cancer.

Data analysis

We compared the individual scores with the expert readers in per patient- and per site analyses. For the patient specific analysis in baseline scans, we compared the Ann Arbor staging classifications (clinically less relevant in NHL, but it was used to have an estimate of observer agreement). Ann Arbor stage I: one nodal station involved or one lymphoid structure (thymus, spleen, tonsil); Stage II: two or more nodal stations on one side of the diaphragm with or without an extra-nodal site on the same side of the diaphragm; Stage III: lymph node stations on both sides of the diaphragm with or without involvement of the spleen; Stage IV: disseminated disease. In scans obtained during or after therapy we compared whether according to the readers the PET scan had normalized. The per site analysis comprised a comparison of the accuracy in localising suspected (extra-)nodal lesions. In a sensitivity analysis, we compared the results of conservative (i.e. equivocal scores considered to be benign) and sensitive reading (assigning equivocal scores to malignant categories), respectively.

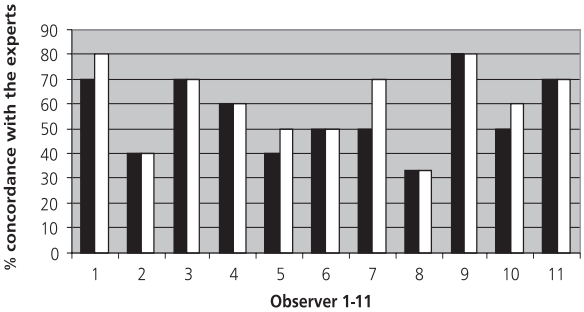
RESULTS

Seven of the 10 pre-therapy scans were classified as Ann Arbor II by the expert readers, and three as Ann Arbor IV. Of the 109 potential classifications (one observer failed to read one scan) obtained from the 11 observers, 66 of 109 (61%) and 61 of 109 (56%) were correct with sensitive and conservative interpretations, respectively (Figure 1). Errors occurred in either direction: overstaging occurred in 24 of 109 (22%) and 30 of 109 (28%) with sensitive and conservative reading, respectively, and understaging in 17% (18 of 109) with either reading system. Independent of the threshold of test positivity, we found a trend towards overstaging with more experienced observers and towards understaging for the less experienced ones; The former group accounted for two-thirds (16 of 24) of the overstaging errors (with conservative reading, vs 18 of 30 with sensitive reading), and the unexperienced for two-thirds (12 of 18) of the understaging errors.

The experts classified 14 of 27 scans obtained during or after therapy as positive for viable tumor. Of the 154 potential tumor positive classifications, 82% was scored accordingly by the observers in conservative reading, vs. 94% in sensitive reading. Thirteen scans were classified as normal by the experts. However, of the 143 potential tumor negative classifications 59% was scored in accordance with the experts (conservative reading, vs 45% for sensitive reading). The experienced observers had slightly better results in patients with remaining active disease (87% correct classifications versus 77% for inexperienced observers with conservative reading, and 97% versus 92% with -sensitive reading). Overstaging in patients without remaining active disease was generally equal for both groups of observers (49% and 51%, respectively) using the conservative assessment strategy and even slightly higher in the experienced group (62% and 49%, respectively) when using the sensitive assessment strategy.

The scans of the 14 patients with active disease harbored a total of 74 involved sites. The experts had no equivocal scores; for the observers the equivocal proportion was higher in the

FIGURE 1 Pre-treatment studies: Ann Arbor classification



Black bars: Sensitive reading: equivocal scores considered malignant.
White bars: Conservative reading: equivocal scores considered benign.

therapy scans (15%) than in the baseline ones (6%). Taken together, the 11 observers scored 554 of the 831 possible tumor sites (67%) in accordance with the expert reading using conservative assessment criteria vs. 73% for sensitive reading. Conservative reading yielded

TABLE 1 Per site analysis of disease classification in 10 baseline scans

Site	Expected correct*	Criteria***	Correct	False positive scores	False negative scores
Nodal					
Neck**	33	C	8 (24%)	32	25 (76%)
		S	11 (33%)	45	22 (67%)
Periclavicular**	77	C	45 (58%)	60	32 (42%)
		S	50 (65%)	62	27 (35%)
Axillary**	33	C	29 (88%)	19	4 (12%)
		S	29 (88%)	21	4 (12%)
Mediastinal/hilar**	184	C	159 (86%)	1	25 (14%)
		S	164 (89%)	1	20 (11%)
Para-aortic/iliac**	44	C	26 (59%)	10	18 (41%)
		S	28 (64%)	12	16 (36%)
Mesenteric	22	C	9 (41%)	5	13 (59%)
		S	12 (55%)	8	10 (45%)
Inguinal**	-	C	-	11	-
		S	-	14	-
Extranodal					
Lung	-	C	-	57	-
		S	-	65	-
Serous	-	C	-	8	-
		S	-	11	-
Liver	11	C	8 (73%)	4	3 (27%)
		S	9 (82%)	6	2 (18%)
Spleen	11	C	8 (73%)	5	3 (27%)
		S	8 (73%)	12	3 (27%)
Bone marrow	44	C	24 (55%)	13	20 (45%)
		S	25 (57%)	18	19 (43%)
All sites	459	C	316 (69%)	225	143 (31%)
		S	336 (73%)	275	123 (37%)

*Expected correct: number of positive sites in case of perfect agreement between observers and experts of the total number of expert-positive sites. **Right and left sided added up and presented together. ***C=Conservative reading: equivocal scores considered benign; S=Sensitive reading: equivocal scores considered malignant.

485 false positive classifications vs. 653 for sensitive reading. A total of 259 classifications was false negative with conservative reading vs. 211 for sensitive reading. Table 1 and 2 shows the geographic variation of these errors. Sources of errors pertained to difficulties in differentiating viable tumour and physiological FDG biodistribution (brown fat, muscle activity, bone marrow uptake after therapy, urinary contamination, leading to false positive readings), to a lesser extent difficulties in anatomical localisation (bulky mediastinal disease

TABLE 2 Per site analysis of disease classification in 27 scans during and after therapy

Site	Expected correct*	Criteria***	Correct	False positive scores	False negative scores
Nodal					
Neck**	22	C	11 (50%)	50	11 (50%)
		S	14 (64%)	65	8 (36%)
Periclavicular**	33	C	25 (76%)	35	8 (24%)
		S	27 (82%)	46	6 (18%)
Axillary**	33	C	27 (82%)	28	6 (18%)
		S	28 (85%)	38	5 (15%)
Mediastinal/ hilar**	108	C	81 (75%)	37	27 (25%)
		S	88 (81%)	55	20 (23%)
Para-aortic/iliacv	55	C	33 (60%)	15	22 (40%)
		S	40 (37%)	13	15 (27%)
Mesenteric	33	C	6 (18%)	13	27 (82%)
		S	7 (21%)	18	26 (79%)
Inguinal**	11	C	4 (36%)	9	7 (64%)
		S	5 (45%)	22	6 (55%)
Extranodal					
Lung	33	C	20 (61%)	36	13 (39%)
		S	28 (85%)	48	5 (15%)
Serous	11	C	9 (82%)	10	2 (18%)
		S	10 (91%)	14	1 (9%)
Liver	-	C	-	7	-
		S	-	10	-
Spleen	-	C	-	9	-
		S	-	19	-
Bone marrow	33	C	22 (67%)	21	11 (33%)
		S	26 (79%)	30	7 (21%)
All sites	372	C	238 (64%)	260	134 (31%)
		S	273 (73%)	378	99 (27%)

interpreted as pulmonary lesions, leading to false positive readings; bowel and mesenterial nodes, ureters and iliac nodes, both leading to false negative readings), and finally to interpretative errors (reading a new lung lesion after therapy as viable tumour while all known sites had disappeared).

With the baseline PET scans false positivity tended to be higher in the neck area and in the lung compared to the other regions. In the remaining scans, the highest concordances with experts were seen in the periclavicular, axillary and mediastinal regions with a tendency towards relatively more errors in the neck regions, and in either direction (ie. false positives and

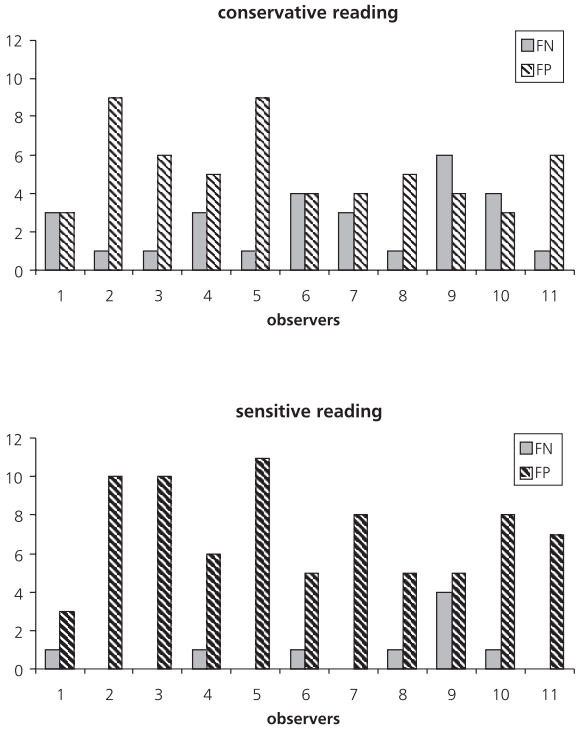
TABLE 3 Analysis of individual observers of lymph node and organ stations in 27 scans used for therapy evaluation.

	Criteria[1]	Obs 1e*	Obs 2e	Obs 3e	Obs 4e	Obs 5e	Obs 6i	Obs 7i	Obs 8i	Obs 9i	Obs 10i	Obs 11i
Neck** n=33	C.	3	3	2	0	3	0	3	1	2	0	1
	S.	3	3	2	1	3	0	3	1	2	2	1
Periclavicular** n=4	C.	4	4	4	1	4	3	4	2	3	0	2
	S.	4	4	4	2	4	3	4	2	3	3	2
Axillary** n=3	C.	3	3	3	2	3	2	2	3	3	2	2
	S.	3	3	3	3	3	2	3	3	3	3	3
Mediastinal/ hilar** n=8	C.	5	8	7	6	7	6	6	6	5	3	6
	S.	6	8	7	6	8	7	7	7	5	6	6
Para-aortic/ iliac** n=6	C.	4	5	4	3	3	4	4	3	4	2	5
	S.	4	5	4	6	3	4	4	3	4	5	5
Inguinal nodes** n=1	C.	1	1	0	0	1	0	1	0	0	1	0
	S.	1	1	0	1	1	0	1	0	0	1	0
Lung n=3	C.	1	2	3	1	2	2	3	2	1	1	2
	S.	3	3	3	2	3	3	3	2	2	2	2
Serous n=1	C.	1	1	1	1	1	0	1	1	1	1	0
	S.	1	1	1	1	1	1	1	1	1	1	0
Liver n=1	C.	1	1	1	0	1	1	1	1	0	0	0
	S.	1	1	1	0	1	1	1	1	0	1	0
Mesenteric n=4	C.	1	0	0	1	2	1	0	2	1	2	0
	S.	1	0	0	1	2	1	0	2	1	3	0
Bone marrow n=4	C.	2	4	4	4	3	3	3	3	2	3	3
	S.	3	4	4	4	3	3	4	3	2	4	3

*C=Conservative reading: equivocal scores considered benign; S=Sensitive reading: equivocal scores considered malignant; *Obs1e: individual observer, number randomly assigned, i and e indicating inexperienced or experienced number of the observers. **Right and left sided added up and presented together; [3]n: Total number of affected sides as scored in the 27 patients by the experts

–negatives). False negativity after therapy was seen more often in the mesenterial and inguinal regions, and false positivity more often in the bone marrow (Table 3 and Figure 2). Differences between sensitive and conservative readings and thus equivocal scores, were most prominent in the neck, around the abdominal aorta/iliac vessels and in the lung (Table 3).

FIGURE 2 Numbers of patient-based classifications at variance with expert readings, in the 27 scans used for therapy evaluation



*Obs1e: individual observer, number randomly assigned, i and e indicating inexperienced or experienced number of the observers.
In case of FN=false negative scores the PET scan was incorrectly classified as negative for viable tumour and in case of FP=false positive scores the PET scan was scored as positive by the observers and negative by the experts

DISCUSSION

FDG PET can provide complementary information to conventional procedures such as contrast enhanced CT and bone marrow biopsy, which leads to modification of stage and has impact on management. Impact of PET findings on staging and patient management varies among different reported studies, up to 60% of patients.[9] To our knowledge, this is the first interobserver variation study on PET in lymphoma conducted in a larger group of nuclear medicine physicians. The main findings were a notable observer variation and errors occurring in either direction, albeit leaning to false positivity. Clinical PET experience had less impact than we had expected, and this is why we pooled the data for most of the analyses. We can only speculate about the lack of impact of clinical experience, but suggest

that with lymphoma, clinical feed-back from the haematologist is less likely to affect reading performance since histopathology of the various suspected lesions as verification is not routinely performed, and if so, is not necessarily correct in case of a negative biopsy. This is quite different from the situation in pre-surgical settings like lung cancer. In fact, this was also the reason for using combined expert readings as the reference standard. Agreement was relatively good for mediastinum, hilus, pleura and spleen, but clearly less well for neck and periclavicular area. The latter variability largely results from variable recognition of aspecific uptake in brown fat and muscle[10], and suggests that observer agreement may improve with a baseline PET scan and/or PET-CT.[11] Baseline PET scans should also help to improve interpretation of bone marrow involvement, which can be quite prominent in scans obtained during treatment due to bone marrow stimulation caused by chemo-immunotherapy. False positive findings at the site of residual masses may be seen due to rebound thymic hyperplasia or post therapy inflammatory changes in lymph nodes or mediastinal/ pulmonary tissue, with the latter apparently substantially more frequent following radiation therapy than after chemotherapy or chemoimmunotherapy.[11;12] Overscore concerning lung lesions were relatively frequent due to bulky mediastinal disease, interpreted as pulmonary lesions. Another source of errors were interpretative errors such as inflammatory lung lesions in case of solitary new lung lesions and good regression of all the known tumour sites on baseline scans. Other sources of overscore were Lack of agreement in treatment evaluation scans with a tendency of overstaging could induce overtreatment and might argue for a baseline (pre-therapy)-scan. Besides PET-CT technology, a uniform approach, ie. a set of test positivity criteria to interpret FDG PET is required to improve observer agreement in this setting. We can only speculate whether the availability of more clinical and CT data, would have affected our results. In malignant lymphoma, Fletcher et al. reported important observer variation with CT readings in patients with Hodgkin's disease and they also proposed for standardization of CT reporting.[13] Even though in nuclear medicine and radiology many test results depend on visual detection and interpretation of individual observers, and observer variation is considered to be the Achilles' heel of the trade, there is a lack of observer variation data with multiple readers (as opposed to limited analyses coming with observational, and often single-center, studies).[14]

CONCLUSION

Interobserver variability in FDG PET readings in lymphoma was higher than anticipated. This variability may affect patient management and the results of clinical trials. A standardized

method of reporting is urgently needed to improve consistency. Moreover PET-CT technology might improve the specificity of interpretations. Availability of baseline scans night also contribute to improve PET readings in these patients.

REFERENCE LIST

1. Jhanwar YS, Straus DJ. The role of PET in lymphoma. J.Nucl.Med. 2006;47:1326-1334.
2. Gallamini A, Rigacci L, Merli F et al. The predictive value of positron emission tomography scanning performed after two courses of standard therapy on treatment outcome in advanced stage Hodgkin's disease. Haematologica 2006;91:475-481.
3. Schiepers C, Filmont JE, Czernin J. PET for staging of Hodgkin's disease and non-Hodgkin's lymphoma. Eur.J.Nucl.Med.Mol.Imaging 2003;30 Suppl 1:S82-S88.
4. Naumann R, Beuthien-Baumann B, Reiss A et al. Substantial impact of FDG PET imaging on the therapy decision in patients with early-stage Hodgkin's lymphoma. Br.J.Cancer 2004;90:620-625.
5. Hoskin PJ. PET in lymphoma: what are the oncologist's needs? Eur.J.Nucl.Med.Mol.Imaging 2003;30 Suppl 1:S37-S41.
6. Spaepen K, Stroobants S, Dupont P et al. Early restaging positron emission tomography with (18)F-fluorodeoxyglucose predicts outcome in patients with aggressive non-Hodgkin's lymphoma. Ann.Oncol. 2002;13:1356-1363.
7. Zijlstra JM, Lindauer-van der WG, Hoekstra OS et al. 18F-fluoro-deoxyglucose positron emission tomography for post-treatment evaluation of malignant lymphoma: a systematic review. Haematologica 2006;91:522-529.
8. Zijlstra JM, Hoekstra OS, Raijmakers PG et al. 18FDG positron emission tomography versus 67Ga scintigraphy as prognostic test during chemotherapy for non-Hodgkin's lymphoma. Br.J.Haematol. 2003;123:454-462.
9. Schoder H, Meta J, Yap C et al. Effect of whole-body (18)F-FDG PET imaging on clinical staging and management of patients with malignant lymphoma. J.Nucl.Med. 2001;42:1139-1143.
10. Castellucci P, Nanni C, Farsad M et al. Potential pitfalls of 18F-FDG PET in a large series of patients treated for malignant lymphoma: prevalence and scan interpretation. Nucl.Med. Commun. 2005;26:689-694.
11. Castellucci P, Nanni C, Farsad M et al. Potential pitfalls of 18F-FDG PET in a large series of patients treated for malignant lymphoma: prevalence and scan interpretation. Nucl.Med. Commun. 2005;26:689-694.
12. Spaepen K, Stroobants S, Dupont P et al. [(18)F]FDG PET monitoring of tumour response to chemotherapy: does [(18)F]FDG uptake correlate with the viable tumour cell fraction? Eur.J.Nucl. Med.Mol.Imaging 2003
13. Fletcher BD, Glicksman AS, Gieser P. Interobserver variability in the detection of cervical-thoracic Hodgkin's disease by computed tomography. J.Clin.Oncol. 1999;17:2153-2159.
14. Robinson PJ. Radiology's Achilles' heel: error and variation in the interpretation of the Rontgen image. Br.J.Radiol. 1997;70:1085-1098.